X, My Search for Significance
(Developing Cost Estimating Relationships)

Hello, my name is turbine inlet temperature, but you can call me X. I'm an independent variable by trade, and this is the story of my search for significance.

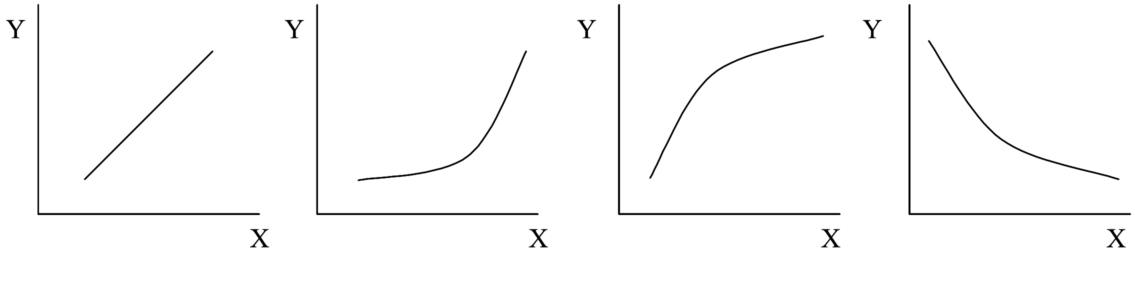
What makes it tick?

I haven't always been an independent variable. When I first started out I was just one of the many characteristics of a jet engine. Then one day a cost and price analyst happened along apparently trying to understand why the cost varied from one jet engine to the next. He began by sitting down with one of the engineers and tried to get an understanding of how a jet engine worked. After that they discussed the different factors that effect or drive the cost of a jet engine. These factors can describe things like the size, performance, or technology of an item. I suppose that if you were interested in trends in wage rates, prices, or something else over time, you could even use time as an independent variable. The analyst made a list of the various factors like weight, thrust, by-pass ratios, and other characteristics like myself because he wanted to develop a parametric or regression model. It seems that the reason the analysts call it "parametric" estimating is that they are trying to explain why the cost varies from one engine to the next by associating the changes in cost with the changes in the engine characteristics or "parameters". I'm told that cost is considered the "dependent" or "Y" variable, and that a dependent variable can be cost, price, hours, pounds, or anything else you want to predict.

Apparently not just anybody can become an independent variable; we had to pass several tests. The analyst didn't want to waste time and money by investigating and collecting data on all the engine characteristics, just the major drivers. He also wanted only "significant" variables. As it turned out, because the by-pass ratios were about the same for all the engines they looked at, the by-pass ratio was not considered to be significant since it didn't vary with the cost. He also seemed particular about knowing with some confidence what the value of the independent variable would be for any engine he needed to estimate. He said, for example, that with software estimating one of the parameters they like to use is size, measured in lines of code or function points. The only problem is that before they can use size to estimate cost, they have to figure some way of estimating the size. The more uncertainty associated with the independent variable, the more uncertainty in being able to predict the value of the dependent variable. And, it goes without saying that an independent variable is no good to you if you can't get the data on it, so some upfront planning is necessary to come up with a reasonable data collection plan.

What's my line?

The next step, for those of us that made the first cut, was to determine in a general sense how we related to cost. Did we make the cost increase or decrease? What "specification" or "function" would best suit us? The analyst said that a function could be linear, curvilinear, or even nonlinear in nature like some of the examples below.



He said that starting out with the right function was important when working with small data sets because one or two data points could easily lead you to the wrong conclusion, so you had better know what you were expecting to see.

In God We Trust (all others must bring data)

They also seemed pretty concerned about how many "analogous" data points or other engines they would be able to find cost and technical data on. The larger the sample of data you have the more the sample looks like the population that it came from and therefore the more confident you would be in any inferences or statements that you would make. Sure, you can run regression on two or three data points, but how much confidence would you have? According to the analyst, when you have a simple regression equation with one independent variable you will lose two "degrees of freedom" because you have estimated the Y-intercept and the slope of the independent variable. So if you have two data points, you have zero degrees of freedom left, with three data points you would have one degree of freedom, and so on. Just like the sample size, the more degrees of freedom the more confident you are in your equation. With each additional independent variable in the equation you lose another degree of freedom. Some analysts[1] suggest that in an ideal situation you should have six to ten degrees of freedom for each independent variable in the equation. Now while that may be desirable, often times we find ourselves with a limited amount of data and have to be a little more realistic. Perhaps a more reasonable rule of thumb would be to count the data points, subtract a degree of freedom for the intercept and one more for each independent variable in the equation, and try to leave yourself with three to five degrees of freedom. You can see how this would affect the number of data points you need to collect and the number of independent variables you could have in an equation. If you only have a few data points you might be better off to find the best one-variable equation rather than using several independent variables together.

Sometimes, good data is hard to come by because it may be: proprietary; collected in a different format; defined differently; not collected at all. Data can also be difficult to compare because things like costs, prices, labor hours, and technology change over time. These things can result in smaller sets of analogous data points. One piece of advice that the analyst offered was to not overly constrain the term "analogous". For instance, you want to determine a reasonable price for a piece of fire fighting equipment on a ship. It might be analogous to equipment used at airports, fire departments, and in industrial settings. Another way that analysts can constrain themselves is to look at something like the F-117 stealth aircraft and say, "There's nothing else like it". Well, maybe there's nothing like the aircraft as a whole, but there may be a number of aircraft using the same landing gear, avionics, engines, and material coatings. The same is true with services. Maybe you don't know what a fair price is for overhauling a truck, but you can find prices for the individual tasks involved in performing an overhaul. Breaking an item or a task down into lower level components is one way of increasing the number of possible analogous data points.

Remember, any regression equation is only as good as the data that was used to create it. It's a lot like painting a room; most of the time is spent in preparation, and if you skip the spackling they won't care how well you rolled the paint on.

[1] Applied Linear Regression Models, Neter, Wasserman, & Kutner, 1989.

85.23567% of all estimates claim a false level of precision

Now, after collecting data on a number of engines, the analyst was going to test myself and some of the other independent variables statistically, using a spreadsheet like Excel or one of the many statistical packages available, to see how well we could explain why the cost varied on jet engines.

One of the tests was to determine how much of the variation in the engine cost I could explain. This was called the R^2 or coefficient of determination, and it is a measure of the *strength* of the relationship between the dependent variable and myself. I was rated on a scale from 0% to 100%. Scoring 0% would have meant that I had nothing to offer, where as scoring 100% would have meant that there wasn't anything about the change in engine cost that I couldn't tell you. The analyst offered two cautions, one, if it sounds too good to be true it probably is; check sample sizes and both the data that was included as well as the data that was not included. Second, the R^2 only represents correlation, not causation; don't go on a fishing trip to find correlated variables, start out with variables that you believe are causal, and then see if there is correlation. Another thing I found out was that as more independent variables are added to the equation the R^2 can only get better, not worse. I was just about to invite some of my friends to join me in the equation so that we could really max-out the R^2 when the analyst warned me about the adjusted R^2. He said that in equations with more than one independent variable he used the adjusted R^2 because it accounted for the loss of the additional degrees of freedom associated with including additional independent variables. Apparently the adjusted R^2 can actually decrease if an independent variable is included that explains little of the variation.

Another test that was run on me was the T test, a measure of the *significance* of the relationship between the dependent variable and myself. The analyst was concerned that factors such as sample size and sampling error could create a false impression of my worth. He assumed that I was not related to engine cost and that it would be up to me to prove otherwise. If I was not significantly related to cost I would have a slope of zero or thereabouts. The larger the slope value, relatively speaking, the more confidence the analyst would have in using me. The T statistic represents the number of standard deviations my slope is from zero, the higher the better. Most applications report a "P" value, which is the probability of my slope being zero. If the P value is .01 then there is only a 1% probability that my slope is zero. Roughly translated, the other 99% is the confidence that you can have in using me. Analysts usually have a pass/fail criterion for using an independent variable based on either an acceptable P value (the level of significance) or an acceptable level of confidence. If I fail the T test in a single independent variable model it means that you would actually prefer to use the average engine cost as your estimate rather than the equation with me. Don't write me off to soon though because even if I don't do well by myself I still might be helpful if paired with one or more other independent variables.

When more than one of us independent variables are used together, our individual T statistics measure the *marginal contribution* we make to the equation. If I fail the T test here you would conclude that you are better off without me in the equation. After all, I cost you a degree of freedom and I wouldn't be giving you anything in return.

One thing you might consider if you decide to use me with another variable is whether the other variable and I are correlated, and how strong is that correlation. Some applications provide a pairwise correlation matrix that can show the correlation between all the variables in a particular model or the correlation between selected variables. This correlation is measured by the R value, which, appropriately enough, is the square root of R^2. The R value can range from a –1 to +1, the plus or minus driven by whether the slope between any two variables is positive or negative. The closer the R value is to one |1| the stronger the relationship between the two variables. The analyst would actually prefer if we were not related since then it would be easier to tell us apart. Some analysts would prefer not to use two independent variables in a model if their R is greater than |.7| even if they have good T statistics. Their concern is that the slopes of the independent variables are less accurate in this case and that they may change from one sample to the next. Other analysts feel that as long as the slopes make sense and have good T statistics there is no harm in using the equation. Given all that, the real issue is that there are two characteristics of, let's say an engine, that are strongly related like the thrust and weight. If the engines in the data set have a thrust to weight ratio of 25% to 30% then we better not use this data, *in any form*, to estimate an engine with a thrust to weight ratio of 45%.

This also brings up a good point about the range of the data which is, stay in the range whenever possible. Take me for example. If the turbine inlet temperatures for the engines in the data set range from 2500° to 3500° then you should not estimate any engine with a value outside this range unless you have the guidance of a technical expert.

Of course, one of the most important considerations is my accuracy; how well do I predict the cost of an engine. In order to answer this question the analyst will look at my standard error (SE) or what is sometimes called the standard error of the estimate (SEE). The SE or SEE is the average or typical estimating error associated with using the equation. If the SE were $50,000, then, *on average*, the estimated engine cost would be off by $50,000 from the actual cost. Sometimes we might be off by a few thousand dollars and at other times by $60,000 or $70,000, but on average we would be off by about $50,000. Now while a SE of $50,000 is a lot of money, it is difficult to put the SE into perspective unless we compare it to the average cost of an engine. If the average engine cost $1,000,000 then the SE of $50,000 would only represent on average a 5% estimating error. This calculation where we divide the SE by the average or mean value is known as the coefficient of variation or CV and it represents the average percent estimating error.

Putting it into Perspective

While there are certainly other important issues to explore like the influence and treatment of outliers, residuals and standardized residuals, and so on, we will leave these for another story. Whether you are building the model yourself or evaluating a model that someone else has developed, the questions are the same. What are the cost drivers? What is the nature of the relationship between the dependent and independent variables? What data points do we include or exclude? What are the statistics and how do we interpret them? Also consider the use of scatter plots because a picture is truly worth a thousand statistics. Keep in mind that when evaluating the results of a parametric model, nothing is more important than the process by which it was built.

Prof Steven Malashevitz
DAU Midwest